Evaluating How Level of Detail of Visual History Affects Process Memory

Eric D. Ragan
Oak Ridge National Laboratory
raganed@ornl.gov

John R. Goodall
Oak Ridge National Laboratory
jgoodall@ornl.gov

Albert Tung
Rio Hondo College
tungay@ornl.gov

ABSTRACT

Visual history tools provide visual representations of the workflow during data analysis tasks. While there is an established need for reviewing analytic processes, and many visual history tools provide visualizations to do so, it is not well known how helpful the tools actually are for process recall. Through a controlled experiment, we evaluated how the presence of a visual history aid and varying levels of visual detail affect process memory. Participants conducted an analysis task using a visual text-document analysis tool. We evaluated their memories of the process both immediately after the analysis and then again one week later. Results showed that even visual history views with reduced data-resolution were effective for aiding process memory. Further, even without inclusion of any data in the visual history aids, the visual cues alone from the final workspace were enough to improve memory of the main themes of analyses.

Author Keywords

Analytic provenance; visual history; process memory

ACM Classification Keywords

H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

INTRODUCTION

Many types of data analysis tasks involve complex investigations with large data sets, open-ended objectives, and iterative hypothesis generation and testing. For example, biological data analysis involves cycles of computational analysis and visualization generation [18], financial analysis requires investigation of vast logs of data to identify suspicious transactions [19], and intelligence analysis includes investigation of large quantities of data from different sources to gather evidence of terrorist activity [21]. For such analyses, the high complexity and potential variability in human analytic processing can make it difficult to remember the steps and rationale that led to the formation of hypotheses, the generation of specific data views, and the realization of conclusions. Process uncertainty can lead to problems, such as when an analyst needs to recall their steps weeks or months after an investigation to review rationale or explain the process to management or other analysts.

To address these issues, researchers have designed provenance tools to help capture and visually represent the history of analytic processes [7, 11]. Provenance tools can serve multiple purposes [12]. While conducting an analysis, workflow logs allow analysts to reference previous stages of an analysis to help keep track of data manipulations or previously explored hypotheses. In complex analyses that consist of multiple sessions or extend over long periods of time, reviewing earlier steps can help an analyst clarify memories of past actions and current goals. In addition to supporting the analytic process itself, provenance tools can be used to help communicate the steps of the process to others. It would be expected that having better memory of process would make it easier to communicate that process, and visual representations of the process might be especially well suited for communication purposes.

While there is an established need for reviewing analytic processes, and many provenance tools provide visualizations to do so, it is not well known how helpful the tools actually are for process recall. Certainly, it would be expected that a complete, fully-detailed record of an analysis would enable a thorough review and lead to a strong understanding of the steps taken and the rationale for the approach. On the other hand, such review could cost a great deal of time and, depending on the size of the data set, could require a large amount of data storage. Lightweight representations of visual history are more practical for portability and quick referencing, but how well do lightweight representations aid recall? And how much data detail is necessary for the visual history to be effective?

We investigated these issues through a controlled experiment in which we evaluated the effectiveness of a visual provenance aid for supporting process memory. Following the methodology outlined in previous work [24], this study demonstrates the development and application of novel methodology for evaluating process memory. In our experiment, participants conducted an analysis and were evaluated on their memories of the process both immediately after the analysis and then again one week

Manuscript published in Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI 2015)

CHI 2015, April 18 - 23, 2015, Seoul, Republic of Korea
http://dx.doi.org/10.1145/2702123.2702376

later. During evaluation sessions, each participant had access to visual history aids based on the state of the analysis workspace at the end of an analysis. By varying the amount of visual detail present in the visual history aids, we assessed the effect of level of detail on process memory.

RELATED WORK

To frame our research, we provide background on visual history tools, the importance of process memory, and the theory of visuospatial context.

Visual History for Analytic Provenance

A variety of visual history tools exist to help researchers and analysts record the process of data analysis. For example, *VisTrails* is a provenance tool that helps track the progression of exploratory visual analytics of scientific data over time [3]. The tool logs the computational steps taken to create different visualizations and generates visualizations of workflow history. Providing analogous functionality, the *GraphTrail* system records and presents analysis pathways taken during exploration of network data [6]. Another tool, *CzSaw*, supports provenance of text document analysis with dependency graphs of entity relationships and visual history of data views throughout an analysis [15]. *ExPlates* is another example that provides visual history during exploratory analysis [14].

A number of researchers have included evaluations of visual history tools. For example, Dunne et al. [6] conducted a three-month field study with archaeologists to understand the practical effectiveness of their *GraphTrail* visualization and to gain insights about how users build visual history maps. The researchers also conducted a qualitative lab study to learn how analysts might use the tool's history tracking functionality. Taking a different approach, Heer et al. [11] analyzed user interaction logs for the *Tableau* visualization software to better understand how users used the undo/redo functionality when working with visual history interfaces.

Such studies are invaluable for understanding how analysts use process-tracking tools and for improving design. To complement qualitative studies, experiments that focus on evaluating the effectiveness of specific design choices can be greatly beneficial. Controlled studies can be used to quantify tool effectiveness or to formally compare specific design options. The challenge is that it is difficult to evaluate the degree to which process history is beneficial, and experimental control often requires reduced ecological realism. One approach is to evaluate the effects of analytic history tools on analysis by considering analysis performance outcomes. For instance, Del Rio and da Silva [5] conducted an evaluation of *Probe-It!*, a provenance visualization tool that shows how maps were created via tree representations that indicate workflow and contributing information sources. Their study found that the majority of participating scientists successfully completed map analysis tasks with the help of the provenance tools, and far fewer successes were observed without the provenance aid. As another example, Groth and Streefkerk [10] studied how different history techniques (no history, undo/redo, and tree history) affected task performance and user confidence for a visual inspection task involving 3D molecules, but they found no significant differences.

Evaluation of Process Memory

While evaluation of analysis outcomes can be useful for determining the effectiveness of provenance visualizations during analysis, real-time support for analysis is just one of the potential benefits of provenance tools. Quality of analysis performance is not indicative of the quality of the *memory* of the analytic process at a later time, which is often necessary for repeating the analysis or communicating the steps to others. The analytic process used by a particular analyst to achieve a given goal will often be a unique approach. Further, many analyses and investigations are exploratory in nature, leading to nonlinear processes involving backtracking and multiple lines of logic. Gotz and Zhou [8] explain that the concept of *insight provenance* involves both the history of steps and their rationale during an analytic process. Analysts should be able to reproduce the logic and approach taken to achieve insights and reach a conclusion. Memory of the analytic process is important for accurate communication, such as during collaboration or for presentation [11].

Some studies have evaluated the effectiveness of visual tools for process memory. Lipford et al. [19] provided insight into this area by comparing recall with and without a visual history tool, although the evaluation was not a controlled comparison and did not focus on specific design components. While controlled evaluations of process memory are limited in the field of visual analytics, studies in workflow support and personal information management are similar (e.g., [4, 20]). Czerwinski and Horvitz [4] conducted research with visual reminder systems to aid workflow memory. In a small user study, the researchers recorded participants for an hour of regular computer work. Later, participants were asked to write down the events that happened during that hour. Participants provided their written memories after 24 hours and then again a month later using both video clips and sets of snapshots from their work sessions. Results indicated that participants preferred snapshots over video as a memory aid.

In a study with a similar type of reminder tool, Park and Furuta [20] evaluated an application that saved continuous screenshots of computer work and allowed users to browse screenshot history. Because the researchers were focusing on supporting task continuity after workflow interruptions, the evaluation consisted of an activity (making travel plans) that was divided over two work sessions separated by one or two days. Participants who used the memory aid were able to resume the task more quickly in the second session as compared to those who did not use the aid.

Visuospatial Context

While the results from studies of workflow-management tools and reminder systems are promising for the feasibility of lightweight visual history tools for data analysis, foundational knowledge is lacking about how to optimize visual design to best support process memory while minimizing data requirements. The issue is becoming increasingly important in the age of big data, in which it is becoming more desirable to support in situ analysis and limit the amount of data saved to permanent storage [1]. In our research, we study whether the visual context of an analysis workspace with reduced data resolution can be effective for eliciting process recall.

During an analysis, the visual workspace of an analysis tool becomes a major component of the environmental context. Through episodic memory, this context can be internalized along with the memory of the activity and data itself [30]. Then, based on the idea that retrieval cues can aid memory of associated content [e.g., 29, 31], perhaps providing the visual context of the workspace will help analysts to recall the process conducted within that context.

The benefit of visual context as a memory aid could become more prominent in the construct of a spatial workspace—such as an analysis environment with a spatial distribution of content (e.g., [2, 6, 18]). Kirsh [17] argued that the spatial organization of items becomes an "integral part of the way we think, plan, and behave." In the context of an office setting, for example, sticky notes in different locations can serve as placeholders that can trigger memories for different actions [16]. In our work, we investigate whether visuospatial placeholders can trigger memories of analysis activities through context alone (i.e., without explicit data). With a spatial distribution, the organization of the workspace can serve as metadata about the process. This notion is further supported by other theory about mapping information to locations. In his spatial indexing model, Pylyshyn [22] discussed referencing information through location. In this way, location acts as an index that can aid recall of associated information.

The value of spatial distribution has also been explored in human-computer interaction. For example, Robertson et al. [26] argued for the benefit of spatial distribution in the *Data Mountain* document-management interface. Studying an intelligence analysis scenario, Andrews et al. [2] discussed how mapping information to physical screen locations helped add a "semantic layer" to the data that was useful for organization and memory. In addition, numerous studies have provided evidence that referring to locations can aid memory of the associated information from those locations [e.g., 13, 23]. Concerning the effects of visual cues on process memory, Ragan et al. [25] studied a procedure memorization task involving the placement of 3D objects. When asked to recall the procedure while viewing the original layout of the workspace, participants demonstrated the best memory of the process when provided with the highest-fidelity visual cues for 3D perception. In our experiment, rather than looking at visual realism, we study how the amount of visual detail influences process memory. Our study focuses on a visuospatial workspace, in which data views and manipulations are mapped to different locations in the workspace.

EXPERIMENT

In this research, we tested whether the use of a visual history aid can improve recall and communication of the analysis process. To accomplish our research goals, we needed a methodology for evaluating process memory. As a proof-of-concept evaluation of process memory, we focus on a simple provenance aid: a visual snapshot of the workspace at the end of an analysis period. A snapshot of visual workspace is appealing because it is lightweight and simple, yet still represents a visual history of user actions and data views. While the simplified tool design sacrifices realism of complex tools and analysis scenarios, it allowed a controlled comparison of different levels of visual detail.

Task

We required participants to perform an analysis activity in order to evaluate analysis memory. The study utilized an intelligence analysis task based on Mini Challenge #1 from the IEEE VAST 2010 Challenge [9], which involves a collection of text records about illegal arms dealing. The data include synthetic records of news articles, government surveillance reports, telephone intercepts, email intercepts, bank transactions, and Internet forum posts. Record lengths vary from single sentences to multiple paragraphs. To accommodate the time constraints of a controlled user study, we simplified the original data set and analysis task by reducing the set to 100 records. Additionally, we shortened some of the longer records.

To focus the analysis task, participants were asked to investigate whether there was a connection between illegal arms dealing and the spread of disease. The task instructions provided participants with the time that the disease was believed to start a pandemic and a list of countries where the disease spread.

Analysis Tool

To complete the analysis task, participants used a custom-built visual text-exploration tool (see top image in Figure 1) on two 24-inch monitors (each 1920 x 1200). Participants used a standard keyboard and mouse for interaction.

The tool presented each text record in a collapsible window. Each window had a header bar that included a date and a phrase to summarize the contents. When collapsed, the full text record was hidden, and only the header bar was visible. Windows could be resized or moved around the workspace. In addition to the windows containing the data records, users could create new editable windows to make notes; these note windows had blue backgrounds to distinguish them from the white windows with the text records.

The tool also allowed users to create connection lines between any two windows. Each connection line included an annotation box with editable text. These lines allowed users to visually indicate relationships between windows.

Users could also use the mouse cursor to select and highlight text. An additional feature that worked with highlighting was the *collapse-to-highlight* function, which worked by hiding all text in a window except for the highlighted text. This functionality was controlled with an additional button in each header bar, and highlight-collapsed windows turned green to make it clear that window was not showing the full, original content.

The tool also supported keywords searching. Search hits were highlighted in pink within the body and header of the window. When a user performed a search, windows containing the search term briefly jiggled to aid visibility.

Experimental Design

The experiment required all participants to first conduct the analysis using the same analysis tool. After the analysis, participants then saw a version of a visual history aid while they were questioned about their findings and processes. The independent variable was the level of visual detail of the visual history aid after the analysis. Four levels of detail were used: *full*, *moderate*, *low*, and *none*. The experiment followed a between-subjects design, so each participant experienced only one of the conditions.

Figure 1 shows partial screenshots as examples of how the tool's workspace might appears in the different conditions. In the *full* detail condition, participants saw the final view of the entire workspace without any modifications. With the *moderate* level of detail, the visual aid included a limited amount of text from the analysis. Only window-header text, highlighted text, and note-window text was visible. In the *low* detail condition, no text was visible. The provenance aid only showed the visual cues from the workspace, including empty windows and colored regions of highlighted content. Finally, in the *none* condition, no visual aid was provided after the analysis; the monitors displayed empty black screens.

Hypotheses

We tested the following hypotheses about how visual history aids and the amount of provided detail can affect process memory:

H1) The *full* detail workspace view will improve process memory. That is, we expect that just the final view of the workspace will help participants remember the steps taken throughout the analysis compared to the absence of any visual aid (the *none* condition).

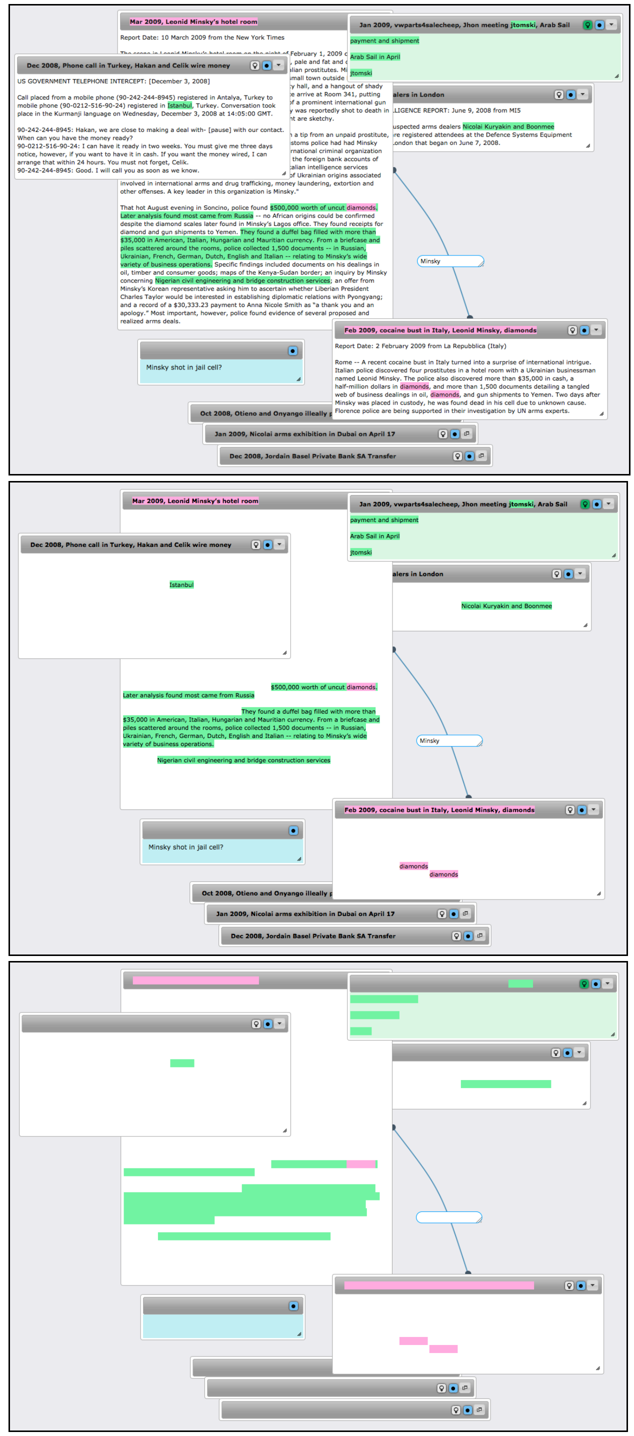

Figure 1. Examples of three of the four levels of visual detail of the visual provenance aid. The top image is example view of *full* detail, the middle image shows *moderate* detail, and the bottom shows *low* detail. The *none* condition is not shown.

H2) A modified view of the workspace with reduced details (*moderate*) will improve process memory. Even if the final view of the state of the workspace only shows a subset of the data that was visible during the analysis, the reduced view will provide advantages over the *none* condition.

H3) Even after removing all data details from the workspace, the visual cues alone will improve process memory. That is, we hypothesize that the visual markup and spatial layout of the *low* detail condition will provide memory benefits as compared to the absence of visual aid.

Participants

A total of 52 volunteers participated in the study, including students and research scientists. Five participants were used for pilot testing to refine the questions needed to elicit descriptions of process memory, and data from four other participants were not included due to vastly different background (research scientists), technical problems, and communication misinterpretations during the procedure.

Data for 43 participants were included in analysis. Eleven participants completed the study in each of the *none*, *low*, and *full* conditions, and ten participants completed the study in the *moderate* condition. Participants were either high school, undergraduate, or graduate students. All participants were interns at a research lab, and the majority studied areas related to science, technology, engineering, and mathematics. Ages ranged from 16 to 30, and the median age was 21. Participants were assigned to conditions to balance gender distribution as well as possible; 14 participants were female (33%).

Procedure

The study design included two sessions separated by one week. In the first session, participants completed a questionnaire about age, gender, and academic program. The experimenter then explained the software to participants using a small practice data set unrelated to the data set used in the analysis task. The experimenter then led a tutorial while participants practiced using the tool. Next, the experimenter explained the intelligence analysis task and objectives. The experimenter also explained how to use think-aloud protocol to provide verbal updates of thoughts, actions, goals, and intentions throughout the analysis. Participants were not told that either of the two sessions would involve recalling the process or findings.

The experimenter then initialized the analysis tool with the data set for the analysis task. All text records started out in the collapsed state, organized in five columns across both monitors. In addition, the tool initialized to include a note box explaining the analysis objective about the spread of disease and illegal arms dealing. Participants were then given 40 minutes to conduct their analyses. Audio and video recordings were used to capture comments and tool usage. The tool automatically logged user actions and periodically saved screenshots of the workspace. If participants were not giving frequent updates (i.e., at least one update every two minutes except when participants were in the middle of reading records), the experimenter prompted for additional updates. Participants were not permitted to take paper notes so that all notes would be contained within the analysis tool.

At the end of the analysis period, the view of the analysis tool was changed to match the appropriate level of detail for the participant's experimental condition, and participants could no longer interact. The experimenter asked a series of questions about participants' findings, followed by questions about analysis strategies. Participants were then asked to provide a step-by-step walkthrough of the steps they took during the analysis, including details about specific actions, hypotheses, dead ends, and goals. Finally, the experimenter asked for further explanation about how tool use, preferred features, and workspace arrangement. This first study session took 65 – 80 minutes.

Participants then returned 6 – 10 days later for the second session, in which they were asked the same questions about their findings and the steps taken through the analysis. Participants sat at the same workstation they used in the first session, and the monitors showed the same visuals corresponding to the appropriate experimental condition. Participants could reference any visuals on the monitors, but interaction was not enabled.

The last component of the second session was a brief step-ordering activity to aid the evaluation of process memory. Between the first and second participant sessions, a member of the research team (not the experimenter) reviewed the record of the analysis (i.e., video, audio, screenshots, and system logs) and employed thematic coding to identify different steps of each participant's analysis. More information about the coding process is described in the following section. From the resulting set of total steps, the researcher distilled the steps down to 10 themes that summarized the main stages or events of the investigation. Then, in the second session, each participant was given a jumbled ordering of the 10 themes and asked to arrange them into the actual order from the investigation. Microsoft PowerPoint was used for this step-ordering activity. Each theme was written on a separate slide, and participants ordered the slides using the *slide sorter* view, which shows miniature views of all slides. The entire second session took approximately 20 minutes.

Evaluation of Process Walkthroughs

Pre-processing was necessary to quantify the results of the verbal process walkthroughs for statistical analysis.

Scoring Process Walkthroughs

While the results of the step-ordering activity provided a simple metric of memory of process order, the primary method of assessing process memory was through scoring the accuracy of participants' verbal walkthroughs of their processes. Two raters separately reviewed the recorded data for each participant's analysis (video, audio, screen shots, and system logs) to understand each actual analysis process. The raters used thematic coding to summarize the steps taken, topics investigated, and strategies used in the intelligence analysis task. For example, if system and video logs showed that a participant conducted multiple searches for terms related to sickness (e.g., *sick*, *disease*, or *infection*), this helped to establish an investigation theme about sickness. As another example, if a participant provided a think-aloud update about a certain arms dealer, and then the corresponding screenshots at that time showed that the participant reviewing several records involving that person, then this would indicate another theme.

Immediately after reviewing and coding each participant's analysis, each rater then reviewed the participant's verbal walkthrough and scored how accurately it described the actual process. Raters gave separate scores for three types of accuracy. First, an *overall score* accounted for a combination of main themes considered, the amount of details given, and the accuracy of the described event order. Second, a *main themes score* was given based on how well the walkthrough covered the main themes or topics covered in the investigation. Lastly, an *ordering score* was given based on the accuracy and clarity that the walkthrough accounted for the order of described steps and themes. All ratings were scored on a continuous scale from zero to ten. Each rater scored both walkthroughs for each participant (i.e., the walkthroughs from session one and two).

Because process coding and scoring were based on each researchers' interpretations of the process, rather than on the participant's thoughts, this evaluation method primarily accounts for observable actions. Even though think-aloud updates were encouraged throughout the analyses, the quality varied. Consequently, the memory assessment has limited consideration for rationale, intentions, and thoughts.

Inter-rater Reliability

Because subjective scoring was used to assess accuracy of verbal walkthroughs, we tested for inter-rater reliability of the two raters. Scores were treated as ordinal measures for reliability testing since the meanings of different values on the scale could have varied between raters. For the purposes of comparing experimental conditions, it was important that raters were consistent in assigning high or low ratings for each outcome, but it was not necessary for raters to agree on exact values. That is, we were interested in inter-rater consistency but not inter-rater agreement.

Following Stemler and Tsai [28], we used Spearman correlations to judge inter-rater consistency. For the *overall scores* and *main theme scores*, correlations were significant with $p < 0.001$ for scores from both sessions, with Spearman's ρ values ranging between 0.72 and 0.78. For *sequence scores*, correlations were significant with $p < 0.005$, and Spearman's ρ values were 0.46 and 0.44 for scores from the first and second sessions, respectively. This demonstrates high inter-rater reliability across all scores.

We also tested intraclass correlation (ICC) between raters using two-way mixed averages measures for consistency with fixed raters, following Shrout and Fleiss [27]. The tests for ICC(3, 2) yielded values between 0.62 and 0.70 for *overall* and *main theme scores*, showing high reliability with $p < 0.001$. For *sequence scores*, ICC measures were 0.38 and 0.33 for sessions one and two, respectively, demonstrating acceptable reliability with $p < 0.05$.

RESULTS

To evaluate the effects of level of visual detail on process memory, we analyzed the results from the process walkthroughs and from the step-ordering activity. Additionally, we considered the amount details given in participants' accounts of their findings in the analysis activity. Because the number of days between the first and second session varied (6 – 10 days), we first tested for correlations between each memory metric and the number of days between sessions. There was no evidence of correlation for any of the metrics.

Variance of Tool Usage

Because participants had the freedom to use the tool as they wanted, specifics of the final view of the visual aid provided during assessment sessions differed for participants. Consequently, some participants used the tool in ways that results in final views with relatively large amounts of supplementary additional cues, while the views for others were visually simplistic. For example, some participants moved and clustered many windows and heavily used the highlight functionality, while others chose not to organize content or use highlighting (see Figure 2).

To account for any potential confounds due to unbalanced tool usage across conditions, we informally reviewed the final views that participants created before formally analyzing the results of the experiment. We ranked the relative amount of window clustering, highlighting, notes created, and connections. As expected, usage strategies and composition of final view varied, but the general compositions within each condition were similar. Each condition included: one to three participants who heavily used clustering; one or two who heavily used highlighting; one or two who heavily used connection lines; and two or three who heavily used the note-taking functionality. We found no evidence of unbalanced conditions with regard to visual complexity of the final workspace.

Memory Results from Walkthroughs

The results for all three types of walkthrough scores (*overall*, *main themes*, and *ordering*) met the assumptions for parametric testing. We tested for effects of the level of visual detail on walkthrough scores using one-way independent ANOVA tests. Separate sets of analyses were run for the results from the first and second sessions.

Session 1 Results: Immediate Recall

The test for *overall scores* from the first session (immediately after the analysis activity) yielded a significant main effect with $F(3, 39) = 5.81$, $p = 0.002$, and $\eta_p^2 = 0.31$. A post-hoc Tukey HSD test showed that both the *full* ($M = 7.43$, $SD = 0.70$) and *moderate* ($M = 6.93$, SD

= 1.55) conditions were significantly better than the *none* condition (M = 5.10, SD = 1.86). Effect sizes were large, with Cohen's d = 1.65 between *full* and *none*, and d = 1.06 between *moderate* and *none*. Other pairwise differences were not significantly different. Figure 3 shows mean *overall scores* for both participant sessions (immediate recall and one week later). Error bars in all bar charts show standard error of the mean.

The ANOVA for *main themes scores* from the first session also showed a significant main effect, yielding $F(3, 39) = 5.91$, $p = 0.002$, and $\eta_p^2 = 0.31$. The post-hoc Tukey test showed the *full* (M = 7.95, SD = 0.61) and *low* (M = 7.29, SD = 1.15) conditions were significantly better than *none* (M = 5.69, SD = 1.84). Effect sizes were again large, with d = 1.65 between *full* and *none*, and d = 1.04 between *low* and *none*. The *moderate* condition (M = 7.19, SD = 1.31) was nearly significantly better than *none* with p = 0.056 and d = 0.93. Figure 4 shows mean *main theme scores* for both participant sessions.

The ANOVA for immediate *ordering scores* showed a significant main effect with $F(3, 39) = 3.69$, $p = 0.020$, and $\eta_p^2 = 0.22$, though the post-hoc Tukey analysis found no evidence of significant pairwise differences.

Session 2 Results: One-week Recall

We also tested the results from the second assessment one week after the analysis activity. The effects from the ANOVA for *overall scores* from the second session were the same as in the first session. The test showed $F(3, 39) = 5.14$, $p = 0.004$, and $\eta_p^2 = 0.28$. The post-hoc Tukey test showed that memory scores for the *full* (M = 6.27, SD = 1.89) and *moderate* (M = 6.22, SD = 1.42) conditions were significantly better than the *none* condition (M = 3.87, SD = 1.53). Effect sizes were large with d = 1.40 between *full* and *none* and d = 1.59 between *moderate* and *none*. Other differences were not statistically significant by the post-hoc test. Figure 3 shows mean *overall scores* and standard error.

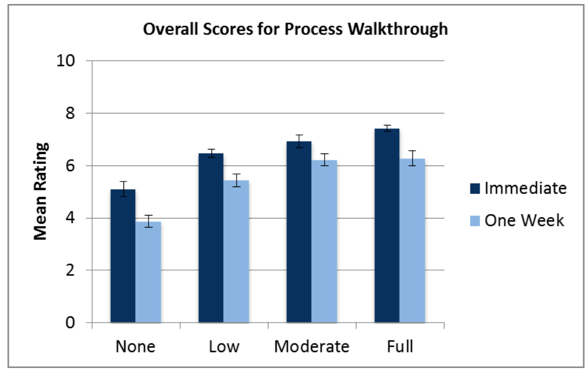

Figure 3. Overall scores for process memory from both verbal walkthroughs. *Full* and *moderate* conditions were significantly better than *none* for both immediate and one-week recall.

The ANOVA on *main themes* results from session two detected a significant main effect with $F(3, 39) = 4.18$, $p = 0.012$, and $\eta_p^2 = 0.24$. Only the *full* (M = 6.87, SD = 1.96) and *moderate* (M = 6.45, SD = 1.29) conditions were significantly better than *none* (M = 4.39, SD = 2.15) by the post-hoc Tukey HSD. Effect sizes were large with d = 1.21 between *full* and *none* and d = 1.15 between *moderate* and *none*. Figure 4 shows mean *main theme scores* visually.

The analysis for effects of visual detail on *ordering scores* one week later showed a significant main effects with $F(3, 39) = 3.65$, $p = 0.021$, and $\eta_p^2 = 0.22$, but post-hoc Tukey testing found no significant pair-wise differences.

Figure 2. Visual history from two different participants in the low-detail condition. The top participant used the tool in a way that provided many visual cues, while the bottom visual provides little more than the original layout.

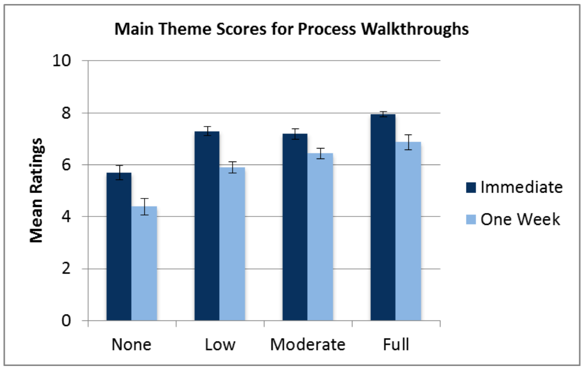

Figure 4. Main topic scores for process memory from both verbal walkthroughs. *Full* and *low* were significantly better than *none* for immediate recall, and *full* and *moderate* were significantly better than *none* after one week.

Memory Results from Step Ordering

The step-ordering activity was only done during the second participant session because it took time to review each participant's analysis and prepare the steps to be sorted. To process the results for formal analysis, we compared each participant's ordering to the correct ordering using Spearman's rank correlation test. We used this correlation statistic as a measure of ordering memory for further analysis, and we tested for effects of level of detail using a one-way independent ANOVA. To account for skewedness (most step-ordering outcomes were relatively high), the results were transformed with the reciprocal function to meet the assumptions of the ANOVA. The test failed to detect a main effect, yielding $F(3, 39) = 0.61$.

As the results were skewed to the left, we believe that the step-ordering activity was generally too easy to be sensitive enough to detect differences in memory. Unlike the process walkthroughs, step ordering allowed recognition of previous steps rather than free recall, and participants were fairly successful at recognizing and ordering steps.

Reported Details in Analysis Solutions

In addition to assessing process memory, we also measured the amount of details that participants provided about their findings from the investigation. To quantify details, we reviewed the verbal accounts of analysis findings and counted the number of names, places, dates, or events that participants mentioned. Means for the four conditions are presented visually in Figure 5.

We tested for effects of visual detail on reported details with one-way independent ANOVAs for reports from the first and second sessions. Results were transformed with *log(x)* to meet the assumptions of parametric analyses, but means and standard deviations are reported untransformed.

The ANOVA on details from the first session found a significant effect with $F(3, 39) = 2.88$, $p = 0.048$, and $\eta_p^2 = 0.18$. A post-hoc Tukey test showed that details mentioned in the *full* condition (M = 11.95, SD = 3.80) were significantly higher than those of the *none* condition (M = 8.09, SD = 3.63). The effect was large with d = 1.04.

The ANOVA for findings details from the second session also found a significant main effect, yielding $F(3, 39) = 3.85$, $p = 0.017$, and $\eta_p^2 = 0.23$. Again, significantly more details were given in *full* (M = 9.36, SD = 3.23) than in the *none* condition (M = 5.32, SD = 1.79) with d = 1.45.

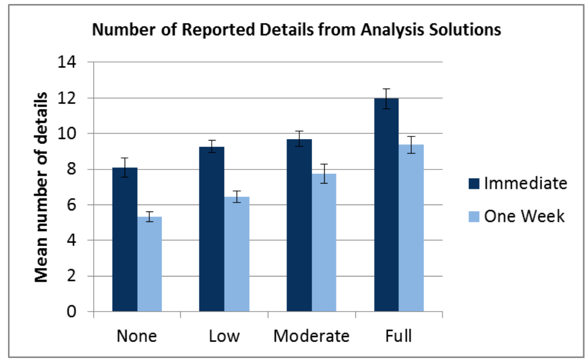

Figure 5. Mean number of reported details from analysis solutions. Significantly more details were given for participants in the *full* condition than in the *none* condition for both immediate and one-week reports.

DISCUSSION

The *overall* scores and *main theme* scores from the process walkthroughs provide strong evidence of the value of a visual history aid for process memory. The results demonstrating the benefit of the *full* detail condition over the *none* condition support hypothesis H1. That is, even without explicitly capturing individual steps throughout the analysis, the final view of a spatial visual workspace significantly improved memory of an analysis process. This finding is interesting because the final view provides only a single static view of the workspace from one point of the analysis (the final point). Yet, as simple as it is, the final view alone was enough to help participants to remember the steps they took throughout the investigation.

Even more interesting are the effects of varying level of visual detail on process memory. The results for the primary memory metrics showed a consistent trend with more visual detail supporting better memory (see Figures 3, 4, and 5). Providing evidence for H2, the *moderate* condition with reduced visible data was significantly better than having no memory aid for process memory (specifically, for overall scores from both sessions and main theme scores from the second session). Presence of headers, highlighted content, and notes was enough to prove beneficial for process memory. This is an important result for the design of provenance tools, as lowering data storage costs is a goal for many types of data analysis. The findings clearly show that reduced data-resolution provenance tools can be effective for helping process memory. Average memory results were similar for the *moderate* and *full* levels of detail, and no post-hoc tests found the *moderate* condition to be significantly worse than the *full* condition; however, scores were still lower in the *moderate* condition. More work is needed to understand how to maximize the benefits of provenance tools while reducing stored details.

What is even more interesting for the discussion of detail reduction—and is perhaps the most noteworthy finding from the study—is the finding that visual cues alone from the final workspace were sometimes enough to improve process memory. This result supports H3. The advantages of the *low-detail* condition over the *none* condition can be seen in Figure 3 and Figure 4 for both immediate and one-week recall. However, the only statistically significant difference between *low* and *none* was for the *main theme scores* for immediate process memory. Because the *low-detail* condition included only visual cues and no textual details, it makes sense that the effect would be stronger for *main themes score*, which did not account for details, than for the *overall score*, which did consider details.

The significant benefit of the low-detail view might be surprising because the condition improved memory of the analysis process without showing any text. Consider the visual aid for the low-detail condition shown in the bottom image of Figure 1. The condition provided only the sizes and positions of windows, the presence of lines, and colored patches where highlighted text had been. Further, having all these types of visual information together in the final view was really the best-case scenario. Refer to the examples from participants in the *low* condition in Figure 2, with the top view showing one of the highest amounts of visual context in the condition and the bottom view showing one of the lowest. Since participants had complete freedom to use the tool however they wanted, most participants did not use all of the functionality. Consequently, the final views contained limited visual information. Despite the limited visual information, the *low* condition provided significant benefits for memory of main themes of analysis throughout a 40-minute session.

Because participants used the tool in different ways, the results of this study can say little about which specific visual cues were most helpful. It is possible that cue effectiveness is dependent on personal preference; for example, perhaps certain users might find more value in positions than from highlighting. More work is needed to understand the types of visual cues that are most helpful for providing memory benefits with minimal detail. However, because the visual aid was based on a spatial distribution, the significant effects of visual cues alone do provide some evidence of the effectiveness of spatially distributed content supporting memory. The findings generally agree with previous studies that found spatial information to be beneficial for remembering content at locations [e.g., 2, 13, 23].

Overall, our study suggests that even extremely simple, low fidelity snapshots that are easy to capture and share can be helpful to analysts in remembering and describing their analysis processes. Even low-resolution visual cues were enough to improve memory of main themes of the analysis, which is promising for the future of reduced-resolution visual history tools. However, to achieve the experimental control needed to detect these effects, the available functionality and visual cues in the study were limited. More work is needed to understand how the effects translate to more complex tools and scenarios.

CONCLUSION

Achieving process memory gains with limited visual and data resolution is important for limiting data storage and allowing portability of visual history tools to smaller devices. We conducted a controlled experiment of how the level of detail available in a visual history aid affects the memory and communication of the analysis process. Level of detail significantly affected both immediate and delayed memories, with more visual detail leading to better recall. Even visual history views with greatly reduced data-resolution were effective for aiding memory, though they were not as effective as the full-detail views. Compared to a complete absence of any visual aid, low-detail visual history views with no visible text significantly improved memory of main themes and stages of the analyses.

These results are promising for the feasibility of lightweight visual history tools to aid analytic provenance. In addition, the study serves as an example of how to evaluate memory of analysis processes, and the results show that the methods successfully detected memory effects. More research is needed to better understand how to take advantage of different types of visual cues to optimize memory benefits while with low-resolution data views.

ACKNOWLEDGMENTS

This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. This research was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Community College Internships program.

REFERENCES

1. Ahern, S. *Scientific discovery at the exascale*. Houston, TX, 2011.
2. Andrews, C., Endert, A. and North, C. Space to think: large high-resolution displays for sensemaking. *Proc. SIGCHI Conference on Human Factors in Computing Systems* (2010), 55-64.
3. Callahan, S. P., Freire, J., Santos, E., Scheidegger, C.

E., Silva, C. T. and Vo, H. T. VisTrails: visualization meets data management. *Proc. ACM SIGMOD international conference on management of data* (2006), 745-747.
4. Czerwinski, M. and Horvitz, E. An investigation of memory for daily computing events. *People and Computers XVI-Memorable Yet Invisible* (2002), 229-245.
5. Del Rio, N. and da Silva, P. P. *Identifying and explaining map imperfections through knowledge provenance visualization*. University of Texas at El Paso, 2007.
6. Dunne, C., Henry Riche, N., Lee, B., Metoyer, R. and Robertson, G. GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. *Proc. SIGCHI Conference on Human Factors in Computing Systems* (2012), 1663-1672.
7. Freire, J., Koop, D., Santos, E. and Silva, C. T. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10, 3 (2008), 11-21.
8. Gotz, D. and Zhou, M. X. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8, 1 (2009), 42-55.
9. Grinstein, G., Konecni, S., Plaisant, C., Scholtz, J. and Whiting, M. VAST 2010 challenge: arms dealings and pandemics. *IEEE Symposium on Visual Analytics Science and Technology (VAST),* (2010), 263-264.
10. Groth, D. P. and Streefkerk, K. Provenance and annotation for visual exploration systems. *IEEE Transactions on Visualization and Computer Graphics*, 12, 6 (2006), 1500-1510.
11. Heer, J., Mackinlay, J., Stolte, C. and Agrawala, M. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE Trans. on Visualization and Computer Graphics*, 14, 6 (2008), 1189-1196.
12. Heer, J. and Shneiderman, B. Interactive dynamics for visual analysis. *Queue*, 10, 2 (2012), 30.
13. Hess, S. M., Detweiler, M. C. and Ellis, R. D. The Utility of Display Space in Keeping Track of Rapidly Changing Information. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 41, 2 (1999), 257–281.
14. Javed, W. and Elmqvist, N. ExPlates: spatializing interactive analysis to scaffold visual exploration. *Computer Graphics Forum*, 32, 3pt4 (2013), 441-450.
15. Kadivar, N., Chen, V., Dunsmuir, D., Lee, E., Qian, C., Dill, J., Shaw, C. and Woodbury, R. Capturing and supporting the analysis process. *IEEE Symposium on Visual Analytics Science and Technology* (2009), 131-138.
16. Kirsh, D. The context of work. *Human–Computer Interaction*, 16, 2-4 (2001), 305-322.
17. Kirsh, D. The intelligent use of space. *Artificial intelligence*, 73, 1-2 (1995), 31–68.
18. Li, G., Bragdon, A. C., Pan, Z., Zhang, M., Swartz, S. M., Laidlaw, D. H., Zhang, C., Liu, H. and Chen, J. VisBubbles: a workflow-driven framework for scientific data analysis of time-varying biological datasets. *ACM SIGGRAPH Asia Posters.* (2011).
19. Lipford, H. R., Stukes, F., Dou, W., Hawkins, M. E. and Chang, R. Helping users recall their reasoning process. *IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2010), 187-194.
20. Park, Y. and Furuta, R. Keeping narratives of a desktop to enhance continuity of on-going tasks. *Proc. of the 8th ACM/IEEE-CS joint conference on Digital libraries* (2008), 393-396.
21. Pirolli, P. and Card, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proc. International Conference on Intelligence Analysis*, 5 (2005).
22. Pylyshyn, Z. The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, 32, 1 (1989), 65–97.
23. Ragan, E. D., Bowman, D. A. and Huber, K. J. Supporting Cognitive Processing with Spatial Information Presentations in Virtual Environments. *Virtual Reality*, 15, 2-3 (2012), 301-314.
24. Ragan, E. D. and Goodall, J. R. Evaluation Methodology for Comparing Memory and Communication of Analytic Processes in Visual Analytics. *Proceedings of BELIV Workshop* (2014).
25. Ragan, E. D., Sowndararajan, A., Kopper, R. and Bowman, D. A. The Effects of Higher Levels of Immersion on Procedure Memorization Performance and Implications for Educational Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 19, 6 (2010), 527-543.
26. Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D. and Van Dantzich, M. Data mountain: using spatial memory for document management. *Proc. ACM symposium on user interface software and technology* (1998), 153-162.
27. Shrout, P. E. and Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86, 2 (1979), 420-428.
28. Stemler, S. E. and Tsai, J. *Best practices in interrater reliability: Three common approaches*. Sage Publications, 2008.
29. Treisman, A. M. and Gelade, G. A feature-integration theory of attention. *Cognitive psychology*, 12, 1 (1980), 97-136.
30. Tulving, E. What Is Episodic Memory? *Current Directions in Psychological Science*, 2, 3 (1993), 67-70.
31. Tulving, E. and Osler, S. Effectiveness of retrieval cues in memory for words. *Journal of experimental psychology*, 77, 4 (1968), 593–601.